Online Methods

Evaluating the quality of SNP calls

Number of SNP calls and allele frequency: The number of calls and frequency for multi-sample calling should follow relatively closely the neutral expectation for N individuals for small N:

$$\text{Number of polymorphic sites} \approx L \cdot \theta \sum_{i=1}^{2N} 1/i$$

where L is the number of confidently called bases and θ is the population-specific heterozygosity, genome-wide of ~0.8 x 10^{-3} for CEPH individuals (Heng Li, unpublished work). A surplus of variants, especially heterozygous variants for single samples or lower-frequency variants for populations, is a strong indicator of false positives.

dbSNP rate: Most variants are already catalogued in the dbSNP database of human variation. For a single European sample, ~90% of their true variants will appear in dbSNP build 129 (Supplemental Table 5), which will reach ~99% following the completion of the 1000 Genomes Project (Supplemental Figure S1). For population-level SNP calls, the aggregate dbSNP rate for the call set decreases as more rare variants are found, which are less frequently found in dbSNP. Nevertheless, the per sample dbSNP rate should remain consistent across individuals, though. Note that presence in dbSNP is not an absolute confirmation that a variant is true (e.g., see Figure 2 and Figure 4), but since dbSNP build 129 contains 11.6M SNP entries (only 0.4% of all genomic positions), relative differences between call sets with high dbSNP rates can be reasonably interpreted as quality differences.

Non-reference sensitivity and non-reference discrepancy (NRD) rate: For single samples, comparison with non-reference (NR) genotype calls from microarray chips, such as HapMap3 (~1.3-1.5M sites), provides a good initial assessment of variant discovery sensitivity. With sufficient coverage, >99% of non-reference sites can generally be discovered. The non-reference discrepancy (NRD) rate reports the percent of discordant genotype calls at commonly called non-reference sites on the chip, and should reach <1% with sufficient coverage. Mathematical definitions of these terms are:

$$\begin{aligned}
\text{NR-sensitivity}(E, C) &= \frac{|E_{nr} \cap C_{nr}|}{|C_{nr}|} \\
\text{NRD-rate}(E, C) &= \frac{|\{i \in E_{nr} \cup C_{nr} : E_i \neq C_i\}|}{|E_{nr} \cup C_{nr}|} \\
X_i &= \text{No. of non-reference alleles for genotype call } i \text{ in call set } X \\
X_{nr} &= \{i \in X : X_i > 0\} \\
E &= \text{Call set to be evaluated} \\
C &= \text{Call set to be compared to}
\end{aligned}$$

Transition/transversion ratio (Ti/Tv): is a critical metric for assessing the specificity of novel SNP calls. Inter-species comparisons [35] and previous sequencing projects (Supplemental Table 6) agree on a Ti/Tv ratio of ~2.0-2.1 for genome-wide data sets and 3.0-3.3 for exonic variation[36]. The expected values for the Ti/Tv for known and novel variants genome-wide are 2.10 and 2.07, respectively, and in the exome target to be 3.5 and 3.0, respectively. Currently the lower Ti/Tv ratio at novel sites than at known sites is due to a combination of residual false positives lowering the Ti/Tv, a relative deficit of transitions due to sequencing context bias, as well as an apparently higher transition ratio at lower frequency variation. These uncertainties should limit the interpretation of minor differences in Ti/Tv ratios (<0.05), especially across sequencing technologies and data sets.

The Ti/Tv ratio for randomly assigned "variation", such as results from systematic sequencing errors, alignment artifacts, and data processing failures, will be ~0.5 as there are two transversion mutations for each transition. Given an expected Ti/Tv

ratio, as above, and an observed Ti/Tv ratio from a call set, an estimate of the fraction of false positive variants in the call set can be obtained by:

$$FDR_{est} = \frac{TiTv_{observed} - 0.5}{TiTv_{expected} - 0.5}$$

which should be bounded above by 100% (due to Ti/Tv ratios below 0.5) and a minimum FP rate (here assumed to be 0.1%) when the observed Ti/Tv exceeds the expected value.

Local multiple sequence realignment

We developed a local realignment algorithm that provides a consistent alignment among all reads spanning an indel. The algorithm begins by first identifying regions for realignment where 1) at least one read contains an indel, 2) there exists a cluster of mismatching bases or 3) an already known indel segregates at the site (e.g. from dbSNP). At each region, haplotypes are constructed from the reference sequence by incorporating any known indels at the site, indels in reads spanning the site, or from Smith-Waterman [37] alignment of all reads that do not perfectly match the reference sequence. For each haplotype H_i, reads are aligned without gaps to H_i and scored according to:

$$\begin{aligned} L(R_j|H_i) &= \prod_k L(R_{j,k}|H_{j,i}) \\ L(R_{j,k}|H_{j,i}) &= \begin{cases} 1-\epsilon_{j,k} \approx 1 & R_{j,k} = H_{j,i}, \\ \epsilon_{j,k} & R_{j,k} \neq H_{j,i}. \end{cases} \end{aligned}$$

$$L(H_i) = \prod_j L(R_j|H_i)$$

where R_j is the *j*th read, *k* is the offset in the gapless alignment of R_j and H_i, and $\varepsilon_{j,k}$ is the error rate corresponding to the declared quality score for the k^{th} base of read R_j. The haplotype H_i that maximizes $L(H_i)$ is selected as the best alternative haplotype. Next, all reads are realigned against just the best haplotype H_i and the reference (H_0), and each read R_j is assigned to H_i or H_0 whichever maximizes $L(R_j|H)$. The reads are realigned if the log odds ratio of the two-haplotype model is better than the single reference haplotype by at least 5 log units:

$$\frac{L(H_0, H_i)}{L(H_0)} = \frac{\prod_j \max\left[L(R_j|H_i), L(R_j|H_0)\right]}{\prod_j L(R_j|H_0)}$$

This discretization reflects a trade-off between accuracy and efficient calculation of the full statistical quantities. Note that this algorithm operates on all reads across all individual simultaneously, which ensures consistency in the inferred haplotypes among all individuals, a critical property for reliable indel calling and contrastive analyses such as somatic SNP and indel calling. The realigned reads are written to a SAM/BAM file for further analysis. The reads around a homozygous deletion, before and after local realignment, for GA reads from the 1000 Genomes Project and HiSeq, are shown in Figure 2.

Base quality score recalibration

We developed a base quality recalibration algorithm that provides empirically accurate base quality scores for each base in every read while also correcting for error covariates like machine cycle and dinucleotide context, as well as supporting platform-specific error covariates like color-space mismatches for SOLiD and flow-cycles for 454[14-16,38,39]. For each lane, the algorithm first tabulates empirical mismatches to the reference at all loci not known to vary in the population (dbSNP

build 129), categorizing the bases by their reported quality score (R), their machine cycle in the read (C), and their dinucleotide context (D). For each category we estimate the empirical quality score:

$$\begin{aligned}
\text{mismatch}(R,C,D) &= \sum_{r\in R}\sum_{c\in C}\sum_{d\in D}\sum_{b_{r,c,d}} b_{r,c,d} \neq b_{\text{ref}} \\
\text{bases}(R,C,D) &= \sum_{r\in R}\sum_{c\in C}\sum_{d\in D} |\{b_{r,c,d}\}| \\
\text{Qempirical}(R,C,D) &= (\text{mismatch}(R,C,D)+1) \,/\, (\text{bases}(R,C,D)+1)
\end{aligned}$$

These covariates are then broken into linearly separable error estimates and the recalibrated quality score Q_{recal} is calculated as:

$$\begin{aligned}
\text{recal}(r,c,d) &= Q_r + \Delta Q + \Delta Q(r) + \Delta\Delta Q(r,c) + \Delta\Delta Q(r,d) \\
\Delta Q &= \text{Qempirical}(R,C,D) - \left(\sum_r \epsilon_r * N_r\right) / \text{bases}(R,C,D) \\
\Delta Q(r) &= \text{Qempirical}(r,C,D) - Q_r - \Delta Q \\
\Delta Q(r,c) &= \text{Qempirical}(r,c,D) - (\Delta Q + \Delta Q(r)) \\
\Delta Q(r,d) &= \text{Qempirical}(r,C,d) - (\Delta Q + \Delta Q(r))
\end{aligned}$$

where each ΔQ and $\Delta\Delta Q$ are the residual differences between empirical mismatch rates and that implied by the reported quality score for all observations conditioning only on Q_r or on both the covariate and Q_r; Q_r is the base's reported quality score and ε_r is its expected error rate; $b_{r,c,d}$ is a base with specific covariate values r,c,d and R,C,D are the sets of all values of reported quality scores, machine cycles, and dinucleotide contexts, respectively. The quality score and covariate distributions for four data sets before and after quality score recalibration are shown in Figure 3.

Multi-sample SNP calling

We apply a Bayesian algorithm for variant discovery and genotyping that simultaneously estimates the probability that two alleles A, the reference allele, and B, the alternative allele, are segregating in a sample of N individuals and the likelihoods for each of the AA, AB, and BB genotypes for each of individual. Given D_i aligned bases at a specific genomic position for individual i, we estimate the genotype likelihoods GT_i of observing the D_i bases for each of AA, AB, and BB genotypes according to the following equation:

$$\begin{aligned}
\Pr\{D_i|GT_i\} &= \prod_j \Pr\{D_{i,j}|GT_i\} \\
\Pr\{D_{i,j}|GT_i = AB\} &= \left(\Pr\{D_{i,j}|A\} + \Pr\{D_{i,j}|B\}\right)/2 \\
\Pr\{D_{i,j}|B\} &= \begin{cases} 1-\epsilon_{i,j} & D_{i,j} = B, \\ \epsilon_{i,j} \cdot \Pr\{B \text{ is true}|D_{i,j} \text{ is miscalled}\} & \text{otherwise.} \end{cases}
\end{aligned}$$

where $\Pr\{D_{i,j} \mid GT_i\}$ is the probability of observing base $D_{i,j}$ under the hypothesized genotype GT_i; $\Pr\{D_{i,j} \mid B\}$, and also $\Pr\{D_{i,j} \mid A\}$, is the probability of observing base $D_{i,j}$ given that the true base is B; $\varepsilon_{i,j}$ is the probability of a base miscall given the quality score of base $D_{i,j}$; and $\Pr\{B \text{ is true} \mid D_{i,j} \text{ is miscalled}\}$ is the probability of B_{true} being the true chromosomal base given that b is a miscall (Supplemental Table 7). As these are raw likelihoods no prior probabilities are applied.

Let us define $q_i = \{0,1,2\}$ as the number of alternate B alleles carried by individual i, so that $q = \sum_i^N q_i$ is the number of chromosomes carrying the B allele among all individuals. We estimate the probability that q = X as:

$$
\begin{aligned}
\Pr\{q = X|D\} &= \frac{\Pr\{q = X\}\Pr\{D|q = X\}}{\sum_Y \Pr\{D|q = Y\}} \\
\Pr\{q = X\} &= \begin{cases} \theta/X & X > 0 \\ 1 - \theta\sum_{i=1}^{2N} 1/i & \text{otherwise} \end{cases} \\
\Pr\{D|q = X\} &= \sum_{GT\in\Gamma}\prod_i^N \Pr\{D_i|GT_i\} \\
\Gamma &= \{GT \text{ where } \sum_i q_i = X\}
\end{aligned}
$$

where Γ is the set of all genotype assignments for the N individuals that contain exactly q = X B alleles, Pr{q = X} is the infinite-sites neutral expectation to observe X alternative alleles in 2N chromosomes with heterozygosity of θ, and GT_i and D_i are the *i*th individual's genotype and NGS reads, respectively. The sum over Γ involves potentially evaluating 3^N combinations but can be approximated by a heuristic algorithm like Expectation-Maximization (EM) via the introduction of a Hardy-Weinberg equilibrium assumption, using a greedy combinatorial search algorithm (Suppl. Mats), or using an exact summation (Heng Li, unpublished results). This algorithm emits the probability of a variant segregating at the site at some frequency:

$$\text{QUAL} = -10 \cdot \log_{10}\left[\Pr\{q = 0|D\}\right]$$

represented conventionally by the Phred-scaled confidence, as well as the genotype assignments at the q* value that maximizes Pr{q | D}. Only sites with QUAL > Q50, for deep coverage, or Q10, for shallow coverage, respectively, are considered here as potentially variable sites.

Variant Quality Score Recalibration

Given a set of putative variants along with their four error covariates (see Supplementary Materials), variant quality score recalibration employs a Variational Bayes Gaussian mixture model (GMM)[40] to estimate the probability that each variant is a true polymorphism in the samples rather than a sequencer, alignment, or data processing artifact. The set of variants $\{v_i\}$ are treated as an n-dimensional point cloud, each variant v_i positioned by its covariate annotation vector, $\vec{v}$. A mixture of Gaussians is fit to the set of likely true variants, here approximated by the variants already present in HapMap3 (Figure 4a). Following training, this mixture model is used to estimate the probability of each variant call being true (Figure 4b), capturing the intuition that variants with similar characteristics as previously known variants are likely to be real, while those with unusual characteristics are more likely to be machine or data processing artifacts.

Mathematically, we write the probability of a variant's vector of covariate values as the linear superposition of Gaussians:

$$\begin{aligned} \Pr\{v_i|GMM\} &= \sum_{k=1}^{K} \pi_k N(\bar{v}_i|\bar{\mu}_k, \Sigma_k) \\ \Pr\{\bar{\pi}\} &= Dir(\bar{\pi}|\alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{a0-1} \\ \Pr\{\bar{\mu}, \Lambda\} &= N(\bar{\mu}|\bar{m_0}, (\beta_0\Lambda_k)^{-1}) W(\Lambda_k|W_0, \upsilon_0) \end{aligned}$$

where K is the number of Gaussians in the mixture (GMM), and the last two equations are standard conjugate prior distributions over the parameters $\vec{\pi}$, $\vec{\mu}$, and Σ.

We then use an analog of the Expectation-Maximization algorithm [40] to learn the optimal parameters for the clusters using only variant calls at sites present in HapMap3. By restricting training to known polymorphic sites, the resulting GMM

captures the distribution of covariate parameters for true SNPs. Consequently, we estimate the likelihood of each putative variant v_i being true under the learned GMM as:

$$
\begin{aligned}
L(v_i|GMM) &= \Pr\{v_i\} \Pr\{\bar{v}_i|GMM) \\
&= (1 - FP_{\text{singleton}})^{AC} \Pr\{\text{novelty of } v_i\} \sum_{k=1}^{K} \pi_k N(\bar{v}_i|\bar{\mu}_k, \Sigma_k) \\
\Pr\{\text{novelty of } v_i\} &= \begin{cases} 97\% & v_i \text{ is in HapMap3,} \\ 37\% & \text{otherwise.} \end{cases}
\end{aligned}
$$

where Pr{v_i} is the prior expectation that the putative variant v_i is true, $\bar{v}_i$ is the vector of covariate values for v_i, $FP_{singleton}$ is the false positive rate for singletons (50% here), and AC is the number of chromosomes estimated to carry the variant, among all called samples. The prior probability of Pr{v_i} depends on whether it is present in HapMap3 and its frequency in the samples being called, given an estimate of the false positive rate for singletons. This model can be easily extended to include more training data, more prior information and/or more error covariates.

For convenience of presentation and analysis, we partition the raw SNP calls into tranches based on the Ti/Tv ratio of their novel variants. For each desired novel false discovery rate target (FDR_i), tranche$_i$ is defined as:

$$
\begin{aligned}
\text{tranche}_i &= \{SNP_j \text{ where } L(SNP_j|GMM) > T_i\} \\
T_i &= \text{smallest } X \text{ where } \text{titv}(\{SNP_j \text{ is novel} \wedge L(SNP_j|GMM) > X\}) > TiTv_i \\
TiTv_i &= FDR_i * (TiTv_{expected} - 0.5)) + 0.5
\end{aligned}
$$

The first tranche is exceedingly specific but less sensitive, and each subsequent tranche in turn introduces additional true positive calls along with a growing number of false positive calls. More specificity in the learned GMM translates into better-separated tranches, where all true variants have high likelihoods and appear

in the lowest FDR tranches, while all false ones have low likelihoods and are excluded. Downstream applications can select in a principled way more specific or more sensitive call sets or incorporate directly the recalibrated quality scores to avoid entirely the need to analyze only a fixed subset of calls but rather weight individual variant calls by their probability of being real.